Горский Павел Владимирович
кандидат технических наук
заведующий сектором методического и
информационного обеспечения мониторинга РИЭПП.
телефон (495) 917 86 66,
pavel@gorskiy.ru

РЕЙТИНГОВЫЙ МЕТОД КОМПЛЕКСНОЙ АТТЕСТАЦИИ ПРЕПОДАВАТЕЛЕЙ СИСТЕМЫ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

Деятельность преподавателя системы высшего профессионального образования отличается значительным количеством особенностей, которые трудно учесть при проведении комплексной аттестации. Особые трудности возникают, если необходимо получить *одну* итоговую, обобщенную оценку. В этом случае почти везде и почти всегда приходится сталкиваться с оценкой в виде так называемого «рейтинга», который, как правило, представляется в виде взвешенной суммы:

$$R = \sum w_i \; r_i \qquad (1)$$

где r_i – оценка преподавателя по i-му критерию (i от 1 до n); w_i - вес i-го критерия.

Трудность в том, что эта формула **не всегда** дает верный результат! Неискушенного читателя это утверждение обычно приводит в недоумение. Следуют заявления вроде того, что приведенная формула «соответствует здравому смыслу», или «отвечает интуитивному представлению о сравнительном качестве преподавателей» и т. п. Однако, все не так очевидно.

Обратимся, прежде всего, к оценкам. Условно разделим оценки на «объективные» и «экспертные». Эксперта упрощенно можно рассматривать как некий «измерительный прибор». Возникает вопрос о точности, с которой этот «прибор» может измерять. Ответ на этот вопрос дан в классической работе Миллера [1]. Там показано, что эксперт чаще всего не способен различать более 7 неименованных градаций. Однако, сплошь и рядом приходится сталкиваться с ситуацией необоснованного использования шкал размерностью выше 7. Например, очень популярна неименованная 10-балльная шкала.

Нередки попытки получить от эксперта информацию в такой форме, в которой он не может дать ее с достаточной надежностью. Последние результаты исследований по этому вопросу приведены в книге Ларичева О. И. и Мошкович Е. М. [2].

Не менее серьезные проблемы связаны с критериями. Прежде всего, не всегда удается *обосновать* тот набор критериев, который необходим и достаточен для выполнения конкретной аттестации преподавателей. Может показаться, что набор критериев «естественно» возникает в каждой конкретной задаче. Но, увы, это далеко не так.

Еще сложнее обстоит дело с весами критериев. Можно даже сказать, что веса критериев – самое тонкое место в проблеме критериального анализа деятельности преподавателей. Чаще всего веса назначают, исходя из интуитивного представления о сравнительной важности критериев. Однако исследования показывают, что человек (эксперт) не способен *непосредственно* назначать критериям корректные численные веса. Необходимы специальные процедуры получения весов.

Итак, при серьезном рассмотрении выходит, что и оценки по критериям и сами критерии с их весами зачастую имеют весьма сомнительное происхождение. Вне поля критики осталась собственно операция суммирования. Оказывается, что такая операция в данном случае не всегда корректна! В классической книге американских математиков Р. Л. Кини и Х. Райфа [3] строго доказано, что подобная формула корректна только тогда, когда все критерии попарно независимы по предпочтению. Что такое «зависимость» критериев, какие виды зависимости бывают и что из этого следует – все это выходит за рамки данной статьи.

Кроме того, оказывается, что *сумма* оценок основана на следующем неявном постулате: «низкая оценка по одному критерию может быть компенсирована высокой оценкой по другому критерию». Однако, этот постулат верен отнюдь не всегда. Например, пусть качество оператора ввода текстов оценивается двумя критериями: 1) скорость ввода (символов в минуту) и 2) среднее количество ошибок на страницу текста. Очевидно, что ухудшение качества ввода (увеличение количества ошибок) не может быть компенсировано увеличением скорости ввода. Можно даже сказать, что в области оценки преподавателей такая ситуация типична. Скажем, недостаток компетентности не может быть компенсирован повышенным уровнем активности.

В итоге, можно сказать, что ценность подобных «рейтингов» вызывает серьезные сомнения. Более того – они скорее вредны, поскольку дают искаженную картину реальности.

Почему же происходят все эти неприятности? Почему с таким постоянством применяется формула (1)? Ответ очевиден – кадровики, консультанты и прочие представители системы высшего профессионального образования, как правило, не знакомы с корректными методами построения обобщенных оценок.

Причем речь идет не только о численных методах. Понятно, что слово «математический» прежде всего ассоциируется с понятием «число». Однако современная математика научилась оперировать не только числами. С начала 70-х годов прошлого века под влиянием запросов прикладных исследований в технических, медицинских и социально-экономических науках в России активно развивается статистика *объектов нечисловой природы*, известная также как статистика нечисловых

данных или нечисловая статистика. Нельзя не отметить, что в создании этой сравнительно новой области прикладной математической статистики приоритет принадлежит российским ученым. Нетрудно показать, что при оценивании преподавателей применимы как числовые, так и нечисловые (например, ранговые) подходы. Только делать это надо грамотно, подбирая подходящий метод к конкретной ситуации. Поскольку рамки статьи не позволяют рассмотреть все подходы, сосредоточим внимание только на численных методах, а точнее – на одной числовой модели.

Определение терминов

Когда говорят «оценка», подразумевают использование какого-либо инструмента измерения. *Оценить* по существу означает *измерить*. Если мы оцениваем, например, квартиру, то мы пытаемся «измерить» ее в денежных единицах. Что же измеряют при оценивании преподавателя? Условно можно считать, что измеряется некоторое «качество» преподавателя как сотрудника, обучающего студентов. Это качество определяется одним или несколькими критериями. Дадим следующие определения.

Оценивание – процедура получения числа, выражающего «качество» преподавателя. Условимся называть это число *рейтингом* преподавателя.

Объект оценивания – преподаватель вуза.

Субъекты оценивания:

- сам преподаватель (если используется самооценка)
- сослуживцы (того же административного уровня, что и преподаватель)
- руководители преподавателя (зав. кафедрой, декан, и т. д.)
- студенты
- психологи
- специалисты предметной области, знание которой необходимо преподавателю в его работе
- сотрудники кадровой службы вуза

В случае использования компьютерного тестирования можно условно считать субъектом оценивания также автора (авторов) теста. Оценку сослуживцев коллегами и начальниками преподавателя назовем *взаимооценкой*.

Критерии – факторы, показатели, определяющие «качество» преподавателя.

Примеры критериев:

- Степень владения профессиональными навыками.
- Умение работать в коллективе.

ЛПР – «лицо, принимающее решение» (или группа таких лиц). Этой аббревиатурой будем обозначать тех руководителей вуза и кадровых служб, которые будут принимать все ответственные решения, связанные с организацией и проведением работ по оценке преподавателей.

Уточнение задачи оценивания

Прежде всего, зададимся вопросом: сколько рейтингов мы хотим получить? Один интегральный (обобщенный) или несколько частных? Например, можно получить два частных рейтинга: 1) рейтинг профессиональных качеств и 2) рейтинг качеств личности. Стоит ли объединять их в один обобщенный? При этом важен и другой вопрос: можно ли корректно построить обобщенную оценку на основе частных? На второй вопрос ответ есть: **можно.** Область прикладной математики, в которой исследуются и развиваются подобные методы, получила наименование «Decision Science» (в русскоязычной литературе часто переводится как «Поддержка принятия решений»). Что же касается первого вопроса, то, вообще говоря, однозначного ответа на него нет. Все определяется целью оценивания. Для определенности в данной статье условимся считать, что нам нужен один обобщенный рейтинг.

Достоверность оценивания

При обсуждении любого рейтинга неизменно встает вопрос о его достоверности. В этом разделе будет представлен математически корректный способ оценки достоверности рейтинга.

Прежде всего, займемся взимооценкой. Предложим всем преподавателям вуза заполнить следующую анкету.

Таблица 1. Степень знания коллег

Перечислите преподавателей, которых Вы знаете	
Профессиональные качества	**Качества личности**
Знаю очень хорошо:	*Знаю очень хорошо:*
(фамилии)	(фамилии)
...	...
Знаю хорошо:	*Знаю хорошо:*
(фамилии)	(фамилии)
...	...
Знаю удовлетворительно:	*Знаю удовлетворительно:*
(фамилии)	(фамилии)
...	...

На основе таких анкет можно построить базу данных, содержащую степень знания преподавателями друг друга. Будем считать, что степень знания – это безразмерная величина в диапазоне от 0 (полное незнание) до 1 (абсолютно полное знание). Договоримся, что абсолютным знанием не может обладать никто и никогда. Тогда, к примеру, можно предложить такую шкалу:

Таблица 2. Шкала числовых эквивалентов качественных оценок степени знания

Степень знания	**числовой эквивалент**
очень высокая	0,9
высокая	0,7
средняя (невысокая)	0,5

Заметим, что низкая степень знания или отсутствие такового нас не интересуют и поэтому не рассматриваются.

Если мы собираемся использовать взаимооценку, то очевидно, во-первых, что чем выше степень знания субъектов оценивания об объекте, тем выше достоверность оценки.

Далее, пусть некоторого преподавателя X будут оценивать N других преподавателей. Тогда нам понадобится также следующая таблица.

Таблица 3. Оценка степени доверия субъективным мнениям

Субъект оценивания	степень знания объекта X	степень доверия ЛПР
Субъект-1	p_1	t_1
...	...	...
Субъект-N	p_n	t_n

Нетрудно заметить, что в последней таблице появилась новая величина – *степень доверия ЛПР*. Определим ее как безразмерную величину в диапазоне от 0 (отсутствие доверия) до 1 (полное доверие). Будем считать, что эта величина выражает степень доверия ЛПР к степени знания конкретного субъекта о конкретном объекте. Например, в последней таблице величина t_1 выражает степень доверия ЛПР к тому, что степень знания субъектом-1 объекта X равна p_1. Поскольку каждый объект оценивания в данной технологии оценивается несколькими субъектами, то для сведения набора степеней доверия объекта оценки к одной (интегральной) степени доверия нам потребуется соответствующая функция. Обозначим ее как

$$V = \phi\,(p,t)\,, \qquad (2)$$

т. е. аргументами функции являются наборы степеней знания и степеней доверия.

Важно отметить, что функция (2) позволяет не только получить интегральную степень доверия для каждого объекта оценивания, но и рассчитать обобщенную степень доверия для всей работы. Полученную таким образом величину можно рассматривать как **степень достоверности** рейтинга.

У читателя может возникнуть вопрос: для чего вводить подобные усложнения? Нельзя ли, к примеру, обойтись без такой неочевидной величины, как **t**? Оказывается – нельзя. Существуют веские обоснования необходимости всех компонентов этой математической модели (см. работу Руссмана [8]). Все, что может показаться излишним, на самом деле жестко требуется для обеспечения *корректности* оценивания.

Критерии: их шкалы и веса

Ясно, что набор критериев оценки управленцев не может совпадать с набором критериев для оценки рабочих и/или специалистов. Стало быть, для каждого типа объекта оценивания должен быть разработан

перечень критериев оценки объектов этого типа. Технологию формирования наборов критериев оставим за рамками данной статьи. Отметим только, что набор критериев должен быть достаточно стабилен, чтобы имелась возможность сравнивать рейтинги, полученные в разное время. Вообще говоря, однократный рейтинг полезен разве что для принятия решений по сокращению штатов. В остальных случаях интерес представляет именно *динамика* рейтинга.

Также существенно для нас и то, что для каждого критерия, по которому предполагается получать *экспертную оценку*, должна быть разработана шкала. Причем, если нет возможности четкой привязки градаций шкалы к какому-то объективному показателю, число градаций не должно превышать 7. Каждая градация шкалы должна иметь содержательное значение (расшифровку). Приведу пример шкалы для критерия «степень владения профессиональными навыками».

Таблица 4. Значения градаций шкалы в баллах

Значение градации	Балл
очень высокая	5
высокая	4
средняя	3
низкая	2
очень низкая	1

Перейдем к весам критериев. Как отмечалось выше, веса критериев – самое тонкое место в проблеме критериального оценивания. Чаще всего веса назначают, исходя из интуитивного представления о сравнительной важности критериев. Однако исследования показывают, что человек (эксперт, ЛПР) не способен *непосредственно* назначать критериям корректные численные веса. Более того, есть данные, (они еще не опубликованы) которые свидетельствуют о том, что человек не может корректно назначать веса даже на базе нечисловых шкал. В рамках прикладной дисциплины Decision Science разработаны надежные методы получения весов критериев, однако их рассмотрение выходит за рамки данной статьи (см. работу Подиновского [9]).

Об одном методе получения численных весов критериев все же следует упомянуть здесь. Метод, о котором пойдет речь, особенно интересен и полезен в том случае, если задача выявления сравнительной важности критериев ставится как отдельная, самостоятельная проблема. В этом случае, наряду с оценками по критериям, каждый субъект оценивания должен выставить каждому объекту, который он оценивает, некую интегральную оценку. Такая «оценка преподавателя в целом» выставляется отдельно от оценок по критериям и никак не должна на них опираться. Если мы говорим «отличный работник» и не детализируем эту оценку, то можно считать, что мы дали человеку некую интегральную оценку. Если у нас есть набор таких интегральных оценок и набор оценок по критериям, то существует корректный математический метод, позволяющий на основе только этой информации получить веса критериев. Причем в рамках метода можно получить также оценку надежности численных значений весов.

Рейтинг

Представим результаты оценок в виде набора следующих таблиц (таблиц будет столько, сколько субъектов оценивания задействовано).

Таблица оценок объектов субъектом S_k

Таблица 5

Объекты оценивания	K_1	K_2	...	K_m
объект - 1	x_{11}	x_{12}	...	x_{1m}
объект - 2	x_{21}	x_{22}	...	x_{2m}
...	...	...	...	...
объект - n	x_{n1}	x_{n2}	...	x_{nm}

Через $\mathbf{K}_j$ обозначен j-й критерий. Через x_{ij} обозначена оценка i-го объекта по j-му критерию.

Ранее мы договорились получить одну интегральную оценку объекта. В этом случае нам понадобится функция

$$R = f(x,w) . \qquad (3)$$

где R – рейтинг объекта, x – вектор критериальных оценок, w – вектор весов критериев.

Поскольку мы также ввели в рассмотрение понятие «достоверность оценки», то в итоге оценка каждого объекта будет представлена двумя числами: R – рейтингом объекта и V – степенью достоверности этого рейтинга. Будем записывать это следующим образом

$$O = \{R,V\} \qquad (4)$$

Возникает естественный вопрос: каков может быть вид функций (2) и (3)? В начале статьи уже говорилось о том, что функция (2) не может быть аддитивной. То же самое, оказывается, справедливо и для функции (3). Вообще говоря, несложно сформулировать систему условий (аксиом), которым должны удовлетворять указанные функции. Труднее подобрать конкретный вид функций, удовлетворяющий всем условиям.

Необходимые базы данных

Какие данные нужны для реализации предлагаемых методов оценивания? Вот примерный перечень:

1. База данных преподавателей вуза, содержащая следующие сведения:

- стоимость 1 часа работы преподавателя,
- степень знания преподавателями друг друга (отдельно в области профессиональной деятельности и отдельно в области личностных характеристик).

2. База данных о психологах, которых можно привлекать к оценке, содержащая номенклатуру методов тестирования с указанием необходи-

мого времени и стоимости по каждой позиции. В эту же базу нужно поместить данные, характеризующие степень доверия для оценок каждого психолога (см. табл. 3).

3. База экспертов по специальностям, степень владения которыми предполагается тестировать. Формат тот же, что и для базы психологов (вид тестирования, время, стоимость, степень доверия).

4. База данных преподавателей, которых предполагается оценивать.

5. База данных, определяющая для каждого объекта оценивания список оценивающих его субъектов.

Оценка и оптимизация затрат

Совокупность указанных выше баз данных позволяет достаточно точно определить общие затраты S на выполнение всего комплекса работ по оцениванию и достигаемую при этом степень достоверности V полученных результатов. Обратим внимание на то обстоятельство, что первые 3 базы данных являются справочными, тогда как базы данных 4 и 5 собственно определяют конкретный профиль работы, т. е. указывают, кого нужно оценивать и кто должен это делать.

Предположим, что базы данных с 1 по 4 неизменны и перед нами поставлена задача проектирования базы данных № 5. При этом можно решать две важные для практики задачи:

Задача 1

При заданном бюджете S подобрать совокупность *субъектов* оценивания таким образом, чтобы степень достоверности V результата была максимальна.

Задача 2

При заданной степени достоверности V результата подобрать совокупность *субъектов* оценивания таким образом, чтобы бюджет S был минимальным.

Обе задачи являются по сути оптимизационными. Для их решения понадобится набор ограничений, например, следующего вида:

- количество психологов, привлекаемых для оценки определенного набора объектов не должно быть менее 2 (вариант – должно быть больше нуля),
- количество оценок для определенного набора объектов не должно быть менее 5,
- и т. п.

Важно, что планирование и разработка проекта по оценке преподавателя в приведенных выше терминах позволяет обоснованно, так сказать «с открытыми глазами», а не на пальцах или на уровне интуиции (как это часто бывает) говорить о соотношении стоимости и эффективности работ.

Типичные ошибки

По нашему опыту, большинство рейтинговых оценок содержат одни

и те же ошибки. Полезно свести типичные ошибки в один список. Некоторые ошибки были уже описаны в начале статьи. Здесь их описание повторяется.

1. Неверный выбор шкал.

В технологии экспертного оценивания эксперта упрощенно можно рассматривать как некий измерительный прибор. Возникает вопрос о точности, с которой этот прибор может измерять. Ответ на этот вопрос дан в классической работе [1]. Там показано, что эксперт не способен различать более 9 градаций. Можно сказать, что в области обработки экспертных оценок закон Миллера играет роль закона Ома в электротехнике. Однако буквально в последнее время появилась возможность использовать шкалы с большим, чем 9, числом градаций. Но для этого нужно применять метод оптимального шкалирования.

2. Незнание особенностей получения данных от экспертов.

Часто делается попытка получить от эксперта такую информацию и в такой форме, которую он не может дать с достаточной надежностью.

3. Непосредственное назначение численных весов критериям.

Во многих работах показано, что такая процедура некорректна. Есть гипотеза (еще не проверенная), что человек не может правильно назначать критериям веса даже в вербальных шкалах. В работе В. В. Подиновского [9] приведена корректная процедура определения весов критериев на базе экспертизы цены замещения критериальных оценок.

4. Агрегирование критериальных оценок с использованием взвешенной суммы.

Строго доказано [3], что взвешенная сумма критериальных оценок корректна только тогда, когда критерии попарно независимы по предпочтению. Это – формальное ограничение для взвешенной суммы. Но есть и содержательные ограничения (см. следующий пункт).

5. Неудачный выбор способа агрегирования.

Любая аддитивная операция (в том числе и взвешенная сумма) может применяться только тогда, когда выполнено условие взаимной компенсации. В противном случае следует рассмотреть другие способы агрегирования [7]. Вообще говоря, строгий подход к выбору методов агрегирования требует детальной проработки аксиоматики, которая должна формализовать цель агрегирования в конкретном проекте.

6. Отсутствие анализа согласованности оценок экспертов.

Если оценки в большой степени рассогласованы, возникает опасность получить так называемую «среднюю температуру по больнице». В зависимости от количества оценок и вида их распределения следует выбирать адекватные способы анализа согласованности.

Литература

1. *Миллер Г.* Магическое число семь плюс минус два. Инженерная психология. М.: Прогресс, 1964.
2. *Ларичев О. И.*, *Мошкович Е. М.* Качественные методы принятия решений. М.: Физматлит, 1996.
3. *Кини Р. Л.*, *Райфа Х.* Принятие решений при многих критериях: предпочтения и замещения. М.: Радио и связь, 1981.
4. *Миркин Б. Г.* Проблема группового выбора. М.: Наука, 1974.
5. *Глотов В. А.*, *Павельев В. В.* Векторная стратификация. М.: Наука, 1985.
6. *Березовский Б. А. и др.* Многокритериальная оптимизация. Математические аспекты. М.: Наука, 1989.
7. *Гафт М. Г.*, *Подиновский В. В.* О построении решающих правил в задачах принятия решений // Автоматика и телемеханика. № 6. 1981.
8. *Руссман И. Б.*, *Бермант М. А.* О проблеме оценки качества // Экономика и математические методы. № 4. 1978. С. 691–699.
9. *Подиновский В. В.* Количественная важность критериев // Автоматика и телемеханика. № 5. 2000.